Trust from Structure: A Traceable, Verifiable Agentic LLM Architecture, Evaluated on VIX Forecasting*

Omar Bairan†

August 9, 2026

Abstract

We present an auditable architecture for agentic language-model systems in financial analysis, together with a protocol that grades it. Five role-specialized agents run as a two-level ReAct loop. An outer state graph sequences the agents, and an inner Tree-of-Thought search keeps competing volatility scenarios open until deterministic verification removes the ungrounded or out-of-bounds ones. Every tool call, prompt, verdict, and intervention enters one audit trail. Six criteria, fixed before the run, grade the system on a walk-forward that issues predictive distributions for the CBOE VIX at one, five, and twenty trading days. Decision-chain coverage is complete, and every searched VIX forecast reconstructs from the logs alone in all four runs. Learning is not demonstrated. Memory on, memory off, and a shuffled-outcome placebo improve at comparable rates, and classical baselines without memory improve at comparable rates on the same dates, which identifies the shared calendar as the source of the improvement. The issued dispersion is neither informative nor reproducible. Its correlation across two runs of the same configuration is -0.06 in the test segment at every horizon. The index moves more after scheduled macroeconomic releases, while the system widens its dispersion after geopolitical headlines that do not move the index. The audit trail localizes the defect that had concealed these results. Each cycle searches one horizon and projects the others, and the twenty-day cycles leave the five-day dispersion and mean unscaled. The evaluation kept those values on dates that both runs share. The unscaled values produced an apparent five-day reproducibility of 0.72 and an apparent system-side response to macroeconomic data ($p = 0.03$), and both disappear once the projection is corrected.

Keywords: LLM agents, auditability, agent memory, human oversight, volatility index, VIX forecasting, evaluation protocols, model risk

JEL codes: G17, C53, C45, C52

*Working paper, August 9, 2026. Sample: CBOE VIX daily closes and classified market news, June 2024 to June 2026. Every number in the paper is reproducible from the scripts and run databases of the replication package, which is available from the author on request.

†`invariant-sigma.ai`. Email: `obairan@invariant-sigma.ai`.

1 Introduction

An agentic language-model system that takes part in a financial decision raises three questions that precede any question about returns. Can each decision be traced to the evidence, reasoning, and rules that produced it? Can the claims of the system be verified independently of the system? Does the system learn from its own record and from the person who supervises it? Much of the agentic trading literature reports backtest returns against buy-and-hold and leaves these questions open, and recent audits find that such returns rarely survive deployment-grade scrutiny (Ye et al., 2026; Xia et al., 2026; Chen et al., 2025). We read this evidence as a problem of measurement order. A system should establish trust before it makes a performance claim. We treat trust as a property of structure, namely of logging, verification, escalation, and learning, and each of these properties admits direct measurement.

The architecture is therefore the contribution of this paper, and the market experiment is its verification. The paper makes three contributions.

1. **An auditable agentic architecture.** Five role-specialized agents (an Orchestrator, Research and Forecasting with an embedded Critic, a Risk and Compliance gate, Reporting, and a closing memory write) run as a LangGraph state graph over three local Ollama models. Models are bound to roles by family, so the agent that makes a claim is never the only agent that approves it. Every capability is a typed, audited tool call over the Model Context Protocol, every prompt and response is traced, and every gate verdict and intervention is a row in one audit trail. Only deterministic checks, a verbatim citation verifier and hard quantitative constraints, may remove a reasoning branch. We evaluate auditability empirically. We report the decision-chain coverage ratio, the full branch-removal census with attribution, the escalation record, and probe evidence that the hard safety rules bind.

2. **A system that learns from its record and from its supervisor.** Three memory loops close over the audit trail. Realized forecast errors update a calibration prior that the reasoning rubric reads on the next cycle. Reflexion notes written after vetoes, escalations, and misses enter later prompts together with normalized supervisor directives, and conflicting directives trigger escalation. A learned confidence multiplier adapts within hard bands. The loops are strictly causal, and controls that can falsify them measure their effect.

3. **A verification experiment on the output of the instrument.** The system ingests only CBOE VIX data (Whaley, 2000) and volatility-related news. It issues predictive distributions at one, five, and twenty trading days over a walk-forward from June 2024 to June 2026, scored against classical baselines (Corsi, 2009; Fernandes et al., 2014). The experiment verifies that the harness operates end to end. It makes no claim of forecasting performance. Its most informative output is a defect in the construction of the system that only the logs could expose (Section 4.4).

Criteria and verdicts. We fixed six criteria before the system ran. Section 4.3 states each criterion formally.

V1 (verification). Full decision chains reconstruct from the logs, hard-rule violations are zero, and calibration is computed per horizon. Verdict: *pass*. Coverage is 100% (114 of 114 searched VIX forecasts in the primary memory-on run, and 123 of 123, 97 of 97, and 95 of 95 in the other three runs).

V2 (learning). Memory on improves calibration over memory off, and the improvement disappears under shuffled outcomes. Verdict: *null*, with an identified cause. The mean absolute standardized error falls at five days in every arm (slopes -0.010, -0.013, and -0.013 per cycle). Random-walk, AR(1), and HAR baselines have no memory, and their errors fall at comparable rates on the same dates (-0.015, -0.011, and -0.008). The decline is a property of the calendar, and the control arms and the baseline control identify it as such.

V3 (supervisor coupling). Escalations defer to the supervisor, resolutions are recorded and bounded, and no supervisor approval overrides the hard veto. Verdict: *mechanism exercised*. The arms raised and resolved 12, 24, and 9 escalations, all through a scripted supervisor policy, and no human participated. No veto fired organically in any run, so a probe establishes non-overridability.

V4 (dispersion informativeness). The issued dispersion predicts realized index variability. Verdict: *null*. In the test segment the issued dispersion of two runs of the same configuration does not reproduce at any horizon (cross-run correlation -0.06 after correction of the projection defect). BV4 has signals at one day (Spearman 0.48) and twenty days (slope 0.73, HAC $t = 1.9$) that BV1 does not reproduce, and the issued variance level never improves on the classical baselines. Section 4.4 gives the mechanism.

V5 (news-class response). Realized variability and issued dispersion differ across news classes, with a response on both sides for the same class. Verdict: *null*. Five-day index moves are 31% larger after macroeconomic data days (permutation $p = 0.0075$, circular shift $p = 0.022$). After correction of the projection defect, the issued five-day dispersion does not widen on those days (0.205 against 0.183, $p = 0.20$). The system widens instead on geopolitical days at every horizon ($p \leq 0.019$), and the index does not move on those days. No class responds on both sides.

V6 (construction consistency). The issued mean embeds the mean reversion of the index. Verdict: *null*. In five-day cycles the issued mean path is linear in the horizon and cannot reproduce the concave reversion profile, and the issued means carry no detectable dependence on the distance of the index from its long-run mean.

2 Related Work

Agentic trading and its evaluation. Language-model trading systems include the layered-memory trader FinMem (Yu et al., 2024a), the multi-agent hierarchy FinCon with conceptual verbal reinforcement (Yu et al., 2024b), the multimodal FinAgent (Zhang et al.,

2024), the debate-structured TradingAgents (Xiao et al., 2024), and HedgeAgents (Li et al., 2025b), with benchmarks such as InvestorBench (Li et al., 2025a) and StockBench (Chen et al., 2025). This literature evaluates systems on returns. A critical strand documents the limits of that evaluation. Reported alpha collapses under controls (Ye et al., 2026), protocols are rarely time-consistent or reproducible (Xia et al., 2026), and memory contributions are rarely isolated (Zhu et al., 2026). We study the system properties that these critiques presuppose: traceability, verification, and measured learning.

Agent patterns and their governance. The architecture composes ReAct (Yao et al., 2023b), Tree of Thoughts (Yao et al., 2023a), Reflexion (Shinn et al., 2023), and deterministic verifiers in the tradition of Cobbe et al. (2021), orchestrated with LangGraph (LangChain, Inc., 2024) and typed tools over the Model Context Protocol (Anthropic, 2024). The self-reported confidence of language models is unreliable (Xiong et al., 2024; Chen et al., 2024). For this reason the system audits confidence through calibration scoring and does not accept elicited confidence as an input. Look-ahead through pretrained weights is documented (Lopez-Lira et al., 2025; Sarkar and Vafa, 2024; Glasserman and Lin, 2023) and is weaker in smaller models (Didisheim et al., 2025), a property that favors the small local models used here.

Model risk governance. Financial supervision has a vocabulary for the properties we measure, and the agentic literature has largely not adopted it. Supervisory guidance on model risk requires independent validation and effective challenge of any model that informs decisions (Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011). Documentation norms such as model cards (Mitchell et al., 2019) and datasheets (Gebru et al., 2021) specify the records that a model and its data require. Record-keeping and traceability obligations for high-risk systems are now regulatory in the European Union (European Parliament and Council of the European Union, 2024) and appear in the NIST risk management framework (National Institute of Standards and Technology, 2023). An agentic system that cannot reconstruct a decision cannot be validated in this sense, whatever its backtest. The coverage ratio, the branch-removal census, and the escalation record express these requirements as quantities that a reviewer can check.

Volatility index forecasting. The VIX is the standard market measure of expected volatility (Whaley, 2000). Its dynamics are strongly persistent and well fit by heterogeneous autoregressive structures (Corsi, 2009; Fernandes et al., 2014), so simple baselines are difficult to outperform and serve as demanding comparators in a verification experiment. Forecast comparison uses the standard toolkit (Patton, 2011; Diebold and Mariano, 1995; Harvey et al., 1997; Hansen et al., 2011; Politis and Romano, 1994; Benjamini and Hochberg, 1995), and proper scoring rules grade distributional quality (Gneiting and Raftery, 2007).

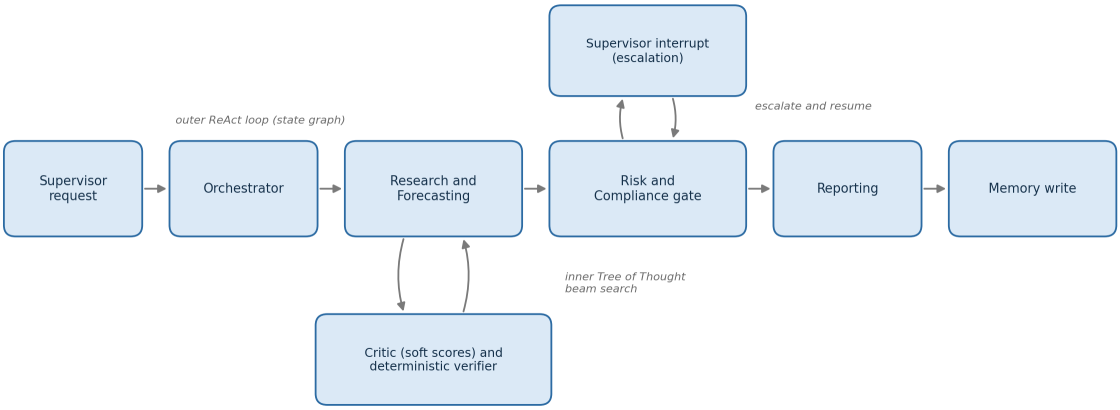

Figure 1: The two-level loop. An Orchestrator, Research and Forecasting with an embedded Critic and a bounded Tree-of-Thought search, a Risk and Compliance gate with a blocking supervisor interrupt, Reporting, and a closing memory write. Every stage writes rows to the five audit layers of Section 3.2.

3 The Architecture as an Audit Instrument

3.1 Separation of duties and deterministic verification

Five role-specialized agents run as a LangGraph state graph over three local Ollama models bound to roles by family. Generation, soft critique, and verification with gating never share a model family, so a judgment does not inherit the errors specific to the model family of the generator. Inside the Research stage a bounded Tree-of-Thought beam search keeps competing scenario branches open. A five-part rubric scores each branch, and the soft score only ranks branches. Branch removal is reserved to the deterministic layer, and the logs confirm that this reservation held in every committed run. The Risk and Compliance gate applies a versioned rule registry and returns approve, veto, or escalate. The gate checks the veto before the escalation branch, so supervisor approval cannot clear a rule violation. An escalation pauses the run at a blocking interrupt. Escalation is part of the policy and does not count as a failure. We evaluate the system on its recognition of the cases in which it should not act alone.

3.2 Five audit layers and their measured use

The system writes five audit layers. (i) `audit_log` records every tool call with its arguments, status, and latency. (ii) `coordination_log` records every inter-agent handoff, tagged by cycle and branch. (iii) `llm_trace` records every prompt and response of every model call, with seed and latency. (iv) Per-cycle reports carry the recommendation, the verified citations, the gate verdict, and the intervention record. (v) A Tree-of-Thought instrumentation store scores every branch of every search.

Coverage. The decision-chain coverage ratio is the share of issued forecasts whose complete chain (evidence retrieval, branch history, verification verdicts, gate outcome, and memory write) reconstructs from the logs alone. Over the searched VIX forecasts the ratio is **100%** (114 of 114 in the primary memory-on run, and 123 of 123, 97 of 97, and 95 of 95 in the memory-off, shuffled, and second memory-on runs). Two disclosures qualify this number. First, the database of the primary run also carries legacy rows from an earlier equity task (123 SPY and 54 QQQ rows, of which 63 are searched forecasts). Nine of those 63 lack complete chains, so pooling them with the 114 VIX forecasts would give a coverage of 94.9% for a population of which 36% belongs to a different task. Second, the ratio is defined over searched forecasts. Each cycle searches one horizon and projects the other two (Section 4.4). Projected rows inherit the chain of their parent cycle by construction, so the 114 searched forecasts correspond to 342 issued rows.

Branch removal. Table 1 reports the census. The architecture claims that only the deterministic layer removes branches, and the claim holds in every run. Across 6,422 VIX branch rows in four runs, no branch was removed by a soft rubric score. The removal rate varies across arms. The shuffled-outcome arm, whose memory holds deliberately mismatched outcomes, removes 6.5% of its branches against 1.5% in the primary run. It is also the arm in which the hard quantitative constraint fires most often (13 times, against twice in the memory-off arm and never in the two memory-on runs). Corrupted memory content leads the generator to propose more ungrounded and more extreme branches, and the deterministic layer removes them. This result is the clearest evidence in the paper that verification has a measurable effect on the output of the pipeline.

Escalation and probes. The memory-on, memory-off, and shuffled arms raised 12, 24, and 9 escalations, each resolved and logged. All other verdicts were autonomous approvals inside the bands (330, 345, and 282). **No veto fired organically in any run.** Non-overridability therefore rests on probes. A restricted-name request is rejected at input and stays vetoed under an explicit supervisor approval, and live order submission stays disarmed throughout. Both conditions are asserted per run, and the full run has zero restricted-name leakage. The hard quantitative constraints run inside every search. The mean-shift cap, which removes a branch whose absolute mean shift exceeds 0.15, removed no branch in the memory-on runs and 15 branches in the control arms. The constraint layer is therefore active during the walk-forward.

Run provenance. Four runs appear in this paper, and every table names its source. BV1 (memory on), BV2 (memory off), and BV3 (shuffled outcomes) form the memory battery and produce Table 3 and Figure 3. BV4 is a second, independent memory-on run over the identical window and snapshots, and it produces Tables 4 and 5 and Figures 4 and 5. We report BV1 and BV4 side by side, because the disagreement between them is a result (Section 4.4). Table 1 covers all four runs.

Table 1: Branch-removal census, VIX searches only, all four runs. Every removal is deterministic. Citation counts removals by the verbatim citation verifier, and Constraint counts removals by the mean-shift cap. Each arm ran 133 searches, and Mean s is the mean search time in seconds.

Run	Branches	Citation	Constraint	Removed	Mean s
Memory on (BV1)	1,592	24	0	1.5%	102.9
Memory off (BV2)	1,615	56	2	3.6%	100.8
Shuffled outcomes (BV3)	1,605	91	13	6.5%	100.7
Memory on (BV4)	1,610	70	0	4.3%	98.3

Table 2: Parameter manifest (excerpt; the full manifest is in the replication package). Every entry is fixed before the segment that uses it.

Parameter	Value	Provenance
Data	CBOE VIX index and news	design
Window	2024-06-01 to 2026-06-30	design
TRAIN	2024-06-03 to 2025-05-30	fixed ex ante
VAL, TEST ends	2025-09-30; 2026-06-30	fixed ex ante
Horizons	1, 5, 20 trading days	design
Cycle cadence	5 trading days (full graph)	design
Beam (K, W, D, τ)	(3, 3, 2, 0.45)	system configuration
Calibration target	cover80 = 0.80 per horizon	design
Memory mode	on, off, shuffled	arm variable
Runs	BV1, BV2, BV3, BV4	Section 3.2
Model digests	7cdf5a01; 7c8c332f; 6577803a	pinned
Memory checkpoints	eb1dc50c… per split, in the package	split boundaries

4 Task, Data, and Notation

4.1 One index, three horizons, dispersion as the object

The system ingests the daily history of the CBOE VIX and a corpus of market news headlines. The local model classifies each headline ex ante, at temperature zero and with a fixed seed, into a six-category taxonomy (monetary policy, macro data, geopolitical, market technical, corporate, other). The classified headlines become point-in-time documents that the verifier can quote verbatim, and dated quantitative context notes derived from the index history complement them. On each decision date the system issues, for each horizon $h \in \{1, 5, 20\}$ trading days, a predictive distribution for the log VIX at $t + h$. The dispersion of that distribution is the empirical object of the paper. Volatility suits an agent that reads text. News moves second moments for days, while its directional content is impounded almost immediately (Ederington and Lee, 1993), and the VIX is the market's own summary of the second moment (Whaley, 2000). Work on the index removes a family of confounds. The design has no portfolio, no exposure policy, no trading rule, and no return to claim.

4.2 Point-in-time discipline

All experiments run over snapshot databases clamped to each decision date. The window splits into TRAIN, VAL, and TEST at the ex ante boundaries of Table 2, and every learned or calibrated quantity uses TRAIN only. The generator and verifier models predate VAL and TEST by at least nine months, so the out-of-sample segments are free of parametric contamination. The pretraining corpus of the generator partially covers TRAIN, and we label TRAIN results accordingly (Lopez-Lira et al., 2025; Didisheim et al., 2025).

4.3 Notation, scores, and tests

Let t index trading days, let V_t be the VIX close, and let $y_t = \ln V_t$. Decisions are taken on the dates $t \in \mathcal{D}$ of a five-trading-day cadence. For $h \in \mathcal{H} = \{1, 5, 20\}$ the realized change is $\Delta_h y_t = y_{t+h} - y_t$. On date t the system issues, for each h, a mean $\mu_{t,h}$ and a dispersion $\sigma_{t,h} > 0$ for $\Delta_h y_t$. All scores treat the issued distribution as Gaussian, $F_{t,h} = \mathcal{N}(\mu_{t,h}, \sigma_{t,h}^2)$, and the standardized error is

$$z_{t,h} = \frac{\Delta_h y_t - \mu_{t,h}}{\sigma_{t,h}}. \tag{1}$$

Calibration and accuracy. Let a segment contain n scored forecasts at horizon h, of which k fall inside the central 80% interval. The coverage is

$$\text{cover80} = \frac{k}{n} = \frac{1}{n} \sum_t \mathbf{1}\{|z_{t,h}| \le z_{0.9}\}, \qquad z_{0.9} = 1.2816, \tag{2}$$

with $H_0\colon \Pr(|z_{t,h}| \le z_{0.9}) = 0.80$. Interval estimates are Jeffreys intervals, the 2.5% and 97.5% quantiles of $\text{Beta}(k + \frac{1}{2}, n - k + \frac{1}{2})$. The probability integral transform $u_{t,h} = \Phi(z_{t,h})$ is uniform under correct calibration. The PIT statistic is the Pearson χ^2 of the counts of $u_{t,h}$ in ten equal bins, referred to a $\chi^2(9)$ distribution under $H_0\colon u_{t,h} \sim U(0,1)$. The continuous ranked probability score of the Gaussian forecast is

$$\text{CRPS}(F_{t,h}, \Delta_h y_t) = \sigma_{t,h} \left[z\big(2\Phi(z) - 1\big) + 2\phi(z) - \frac{1}{\sqrt{\pi}} \right], \qquad z = z_{t,h}, \tag{3}$$

where ϕ and Φ are the standard normal density and distribution function (Gneiting and Raftery, 2007). Point accuracy uses the squared and absolute errors of the implied log level. For a paired comparison of CRPS on common dates, the minimum detectable effect is $\text{MDE} = (z_{0.975} + z_{0.80})\, s_d/\sqrt{n}$, where s_d is the standard deviation of the paired differences, at a two-sided size of 5% and a power of 80%.

Learning curves. Order the scored forecasts of a run and horizon by cycle, $j = 1, \ldots, J$, and let $\bar{a}_j = \frac{1}{8} \sum_{i=j-7}^{j} |z_i|$ be the rolling mean absolute standardized error over eight cycles. The

learning-curve slope b is the OLS coefficient in

$$\bar{a}_j = a + b\, j + e_j. \tag{4}$$

Its t statistic uses a Bartlett-kernel heteroskedasticity and autocorrelation consistent (HAC) variance with eight lags (Newey and West, 1987), under $H_0\colon b = 0$. A significantly negative b is a decline in standardized error. Under correct calibration $\mathbb{E}|z_{t,h}| = \sqrt{2/\pi} \approx 0.80$. A negative slope therefore records shrinking errors relative to the issued width, and it indicates a calibration gain only while $\bar{a}_j$ exceeds 0.80.

Dispersion informativeness. For each horizon we estimate by OLS

$$|\Delta_h y_t| = \alpha + \beta\, \sigma_{t,h} + \varepsilon_{t,h}, \tag{5}$$

with $H_0\colon \beta = 0$ against $H_1\colon \beta > 0$. The HAC variance of $\hat{\beta}$ uses a Bartlett kernel with $\max(1, h/5)$ lags, which counts the overlapping evaluation windows at the five-day cadence. We also report $R^2 = 1 - \sum_t \hat{\varepsilon}^2_{t,h} / \sum_t (|\Delta_h y_t| - \overline{|\Delta_h y|})^2$ and the Spearman rank correlation ρ_S between $\sigma_{t,h}$ and $|\Delta_h y_t|$. The level of the issued variance enters a loss comparison. With the realized variance $RV_{t,h} = \sum_{s=1}^{h} (\Delta_1 y_{t+s-1})^2$ of daily log changes over the horizon, the QLIKE loss of a variance forecast f is (Patton, 2011)

$$L(f, RV_{t,h}) = \frac{RV_{t,h}}{f} - \ln \frac{RV_{t,h}}{f} - 1. \tag{6}$$

The comparators are a TRAIN-constant variance, a rolling 21-day variance, and a HAR model of squared daily changes.

Forecast comparison. For two forecasts with loss differential $d_t = L_{1,t} - L_{2,t}$, the Diebold–Mariano statistic is $\mathrm{DM} = \bar{d} / \hat{V}(\bar{d})^{1/2}$ with a HAC variance. The small-sample correction of Harvey, Leybourne, and Newbold refers

$$\mathrm{DM}^{*} = \mathrm{DM} \left[\frac{n + 1 - 2m + m(m-1)/n}{n} \right]^{1/2} \tag{7}$$

to a $t(n-1)$ distribution, where $m = \max(1, h/5)$ is the overlap of the forecast windows in units of the cadence (Diebold and Mariano, 1995; Harvey et al., 1997). The hypotheses are $H_0\colon \mathbb{E}[d_t] = 0$ and $H_1\colon \mathbb{E}[d_t] \neq 0$. Within each table family we control the false discovery rate at $q = 0.10$ (Benjamini and Hochberg, 1995).

News-class response. Let E_c be the set of event days of class c, defined by a per-class threshold on the daily headline count. The market statistic is the mean of $|\Delta_5 y_t|$ over $t \in E_c$, set against its mean over all days. Its permutation p-value is the share of 2,000 random day sets of

size $|E_c|$ with a mean at least as large. Its circular-shift p-value is the share of 1,000 circular shifts of the class calendar with a mean at least as large, and the shift preserves the clustering of event days. The system statistic is the difference between the mean issued five-day dispersion on class decision dates and on the other decision dates, with a two-sided p-value from 2,000 random relabelings. The conjunction test treats the six classes as one family. Under H_0 the class calendars carry no information on either side. We shift all six calendars by a common random offset 2,000 times, recompute both statistics for every class, and report the share of shifts in which at least one class attains a market statistic and a system statistic both at least as large as the observed pair of macro data.

Construction consistency. Let $g_t = y_t - \bar{y}_t$ be the gap between the log VIX and its point-in-time expanding mean since 1990. An Ornstein–Uhlenbeck process with reversion speed κ and long-run level θ implies

$$\mathbb{E}_t[y_{t+h}] - y_t = (1 - e^{-\kappa h})(\theta - y_t), \tag{8}$$

a concave function of h. We regress the issued mean $\mu_{t,h}$ on g_t, with $H_0\colon \gamma = 0$ against $H_1\colon \gamma < 0$ for the slope γ, and compare the estimate with the benchmark slope of a walk-forward regression of the realized change $\Delta_h g_t$ on g_t.

Criteria. V1 is an accounting criterion with two conditions, a coverage ratio of one and zero hard-rule violations. V2 requires that memory on improve on memory off in CRPS, coverage error, PIT χ^2, or the learning-curve slope, and that the shuffled arm fail to reproduce the improvement. V3 is a mechanism criterion on escalation, recording, and the veto. V4 has two conditions. The slope condition rejects $H_0\colon \beta = 0$ in (5) on a dispersion series that reproduces across runs, and the loss condition compares the issued variance level in (6) with the best classical baseline. V5 requires a class that rejects its null on both sides and survives the conjunction test. V6 rejects $H_0\colon \gamma = 0$ with a slope of the sign and order of the benchmark.

4.4 One search per cycle, two cycle types

The system does not search three horizons. Each cycle runs one full Tree-of-Thought search at a single anchor horizon and derives the other two horizons analytically. Across the four runs the anchor is five days in 347 of 429 complete cycles (93 of 114 in BV1, 99 of 123 in BV2, 78 of 97 in BV3, and 77 of 95 in BV4) and twenty days in the remainder. The twenty-day cycles follow a schedule of about twenty trading days. In BV1 and BV4 every date that carries both cycle types (15 and 16 dates) has its twenty-day cycle issued minutes after its five-day cycle. We recovered the derivation rules from the logs alone. They hold to four decimal places.

- *Five-day cycles.* $\sigma_1 = \sigma_5/\sqrt{5}$, $\sigma_{20} = 2\sigma_5$, and $\mu_h = \mu_5\, h/5$. The dispersion follows

the square-root-of-time rule, and the mean is linear in the horizon. The median of $(\mu_{20} - \mu_1)/(\mu_5 - \mu_1)$ lies between 4.749 and 4.750 in every run, and an exactly linear path gives $19/4 = 4.75$.

- *Twenty-day cycles.* $\sigma_1 = \sigma_{20}/\sqrt{20}$, while $\sigma_5 = \sigma_{20}$ and $\mu_5 = \mu_{20}$ without scaling. The rules of the five-day cycles give $\sigma_5 = \sigma_{20}/2$ and $\mu_5 = \mu_{20}/4$. The one-day mean follows no fixed ratio to μ_{20}.

- *Row selection.* The evaluation harness keeps the last issued row per date and horizon. On a date with both cycle types it therefore keeps the twenty-day cycle, and the five-day forecast of that date carries the unscaled values.

We call this rule the harness selection. The corrected selection keeps the same rows and replaces the unscaled five-day values of twenty-day cycles by $\sigma_5 = \sigma_{20}/2$ and $\mu_5 = \mu_{20}/4$. A robustness selection keeps only the five-day cycles. Tables 3, 4, and 5 and Figures 3 to 5 use the corrected selection, and the text reports the harness values where they differ. The correction changes only the five-day forecasts of twenty-day cycles. Four consequences follow.

The horizons are not three independent tests. The Spearman correlation is invariant to monotone transformations. Under the corrected selection the three dispersions of a date are fixed multiples of one another, $\sigma_1 : \sigma_5 : \sigma_{20} = 1 : \sqrt{5} : 2\sqrt{5}$. A rank-order result at one horizon therefore restates the ranking of a single issued quantity, and V4 is one predictor evaluated against three outcome horizons.

The unscaled values produce the apparent reproducibility. Both runs issue twenty-day cycles on the same calendar. The type of the kept cycle agrees on 68 of 74 common decision dates (92%) and on 25 of the 27 common TEST dates, and on 5 of those 27 dates both runs keep a twenty-day cycle. On those dates the harness five-day dispersion is about twice its consistent value in both runs. Under the harness selection the cross-run Pearson correlation of the issued dispersion on the 27 TEST dates is 0.72 at five days and -0.06 at one and twenty days, and the shared doubled values produce the 0.72. Under the corrected selection the correlation is -0.06 at every horizon, and with five-day cycles only it is -0.15 at every horizon on 25 dates. Over all three segments the corrected correlation is 0.66. *In the test segment the issued dispersion does not reproduce across runs at any horizon.*

The unscaled values reach other results. The same rows enter every five-day evaluation. Three of the seven BV4 decision dates on macroeconomic data days are twenty-day cycle dates, and the correction removes the apparent system-side response of V5 (Section 6). The correction also lowers the five-day CRPS and PIT statistics of the memory arms (Table 3). It leaves every verdict other than V5 unchanged.

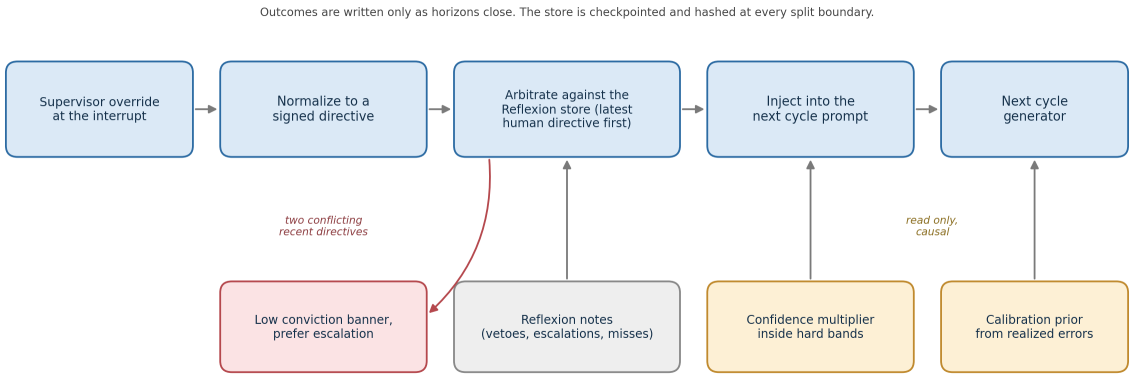

Figure 2: The intervention pathway of memory. A supervisor override is normalized to a signed directive and arbitrated against the stored Reflexion lessons, and the most recent human directive takes precedence over them. Two conflicting recent directives raise a low-conviction banner and a preference for escalation. The calibration prior and the confidence multiplier are read-only inputs that adapt inside hard bands.

V6 fails at the construction level. In five-day cycles the mean path is linear in h and cannot reproduce the concave profile of (8), and in twenty-day cycles the five-day mean equals the twenty-day mean. The regression adds a second result. Against a benchmark slope of -0.180 ($t = -8.9$) at twenty days on the 1990–2024 history, the issued means give $+0.008$ with a 95% interval of $[-0.032, +0.047]$, and the interval excludes the benchmark. The mean profiles come from the projection layer, a deterministic component, and the audit trail localized them there. Without the logs the same evidence would read as a general failure of the agent to represent its object.

5 Memory as a Measured Component

5.1 Three loops, strictly causal

Loop one writes calibration evidence. As each horizon closes during the walk-forward, the realized index value is written against the forecast that predicted it (error, standardized error, and probability integral transform), and the reasoning rubric reads a calibration prior computed from that record on the next cycle. Loop two writes lessons. Reflexion notes (Shinn et al., 2023) written after vetoes, escalations, and the worst calibration misses enter later generator prompts, together with supervisor overrides normalized to signed directives, through the arbitration rule of Figure 2. Loop three adapts confidence. A learned multiplier that summarizes realized hit rates adapts inside hard bands and is isolated per split. The system enforces causality and logs the evidence. Outcomes are written as horizons close and never retrospectively (118 written in run under memory on, 6 under memory off, and 102 permuted under shuffled outcomes), and the

memory store is checkpointed and hashed at the TRAIN and VAL boundaries.

5.2 The measurement program

Arms. Memory on and memory off run the identical walk-forward under identical seeds and snapshots. The off arm uses a neutral prior, no injections, and a multiplier fixed at one. A content control warm-starts from the TRAIN-end checkpoint of the memory-on arm and permutes outcome content while it leaves the machinery intact. This control detects a formatting effect that would otherwise pass as learning, and any learning claim must beat it. The TEST segment holds 36, 37, and 28 five-day decision dates for the on, off, and shuffled arms. The on and off calendars differ by one date, which the log records, and the shuffled arm is shorter because it warm-starts.

A baseline control for the calendar. The decisive measurement for V2 uses classical baselines. We score random-walk, AR(1), and HAR predictive densities on the same decision dates and horizons as the agent, with residual widths estimated on TRAIN, and run the identical learning-curve regression (4). These models have no memory of any kind. If their errors decline at the rate of the agent, no memory mechanism can claim the decline of the agent.

Results. Table 3 and Figure 3 report all three horizons. At five days the rolling mean absolute standardized error declines significantly in every arm (slopes -0.010, -0.013, and -0.013 per cycle; harness selection -0.011, -0.016, and -0.012). At one and twenty days no arm has a detectable trend, and the largest of the six slopes has $|t| = 1.8$. The memoryless baselines behave in the same way. At five days their slopes are -0.015 (random walk, $t = -13.3$), -0.011 (AR(1), $t = -6.0$), and -0.008 (HAR, $t = -5.1$), and at twenty days they are $+0.005$, $+0.001$, and $+0.000$, none distinguishable from zero. **Criterion V2 returns null, and the baseline control identifies the cause directly.** The later part of the window is a regime of smaller forecast errors for every predictor, with or without memory. Figure 3 places the five-day rolling mean below the calibrated value of 0.80 throughout TEST in every arm. The declines therefore record shrinking errors relative to issued widths that already exceed the realized dispersion, in line with the coverage above nominal in Table 3.

The comparison has limited power, and Table 3 quantifies it with the minimum detectable CRPS difference at the realized sample sizes. The observed gap between memory on and memory off is 0.46 of the detectable effect at one day, 0.12 at five days, and 0.15 at twenty days. This single-seed design could not detect a memory effect smaller than about twice the observed one at one day, or smaller than six to nine times the observed one at five and twenty days, and we do not treat the absence of evidence as evidence of absence. The off arm has one asymmetry. Reflection writes still occur in it, because they belong to the audit path, and only the read-back is disabled.

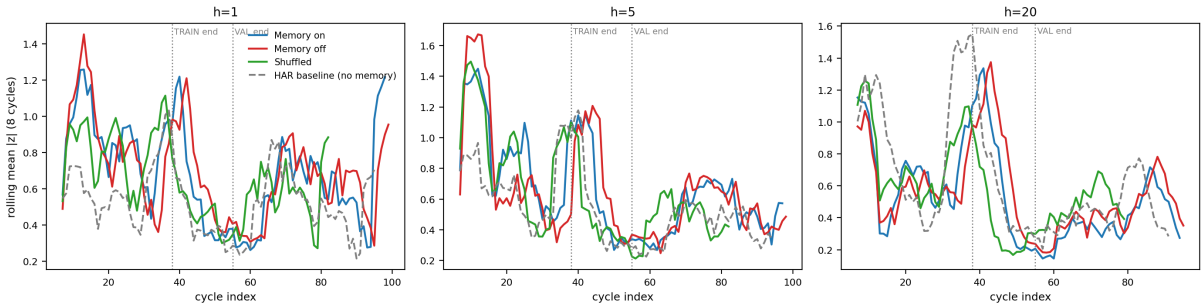

Figure 3: Rolling mean absolute standardized error (mean $|z|$ over eight cycles) against cycle index for the three memory arms and, dashed, a HAR baseline without memory scored on the same dates. Dotted lines mark the split boundaries. The baseline tracks the arms at every horizon, and this co-movement identifies the within-run decline at $h = 5$ as calendar structure. Runs BV1, BV2, BV3, corrected selection (Section 4.4).

6 The Verification Experiment

6.1 Design

The walk-forward runs the full agent graph every five trading days over the snapshot data, and light updates run daily. We report forecast accuracy against baselines for completeness. The paper makes no accuracy claim, and the VIX is persistent enough that simple models are demanding comparators (Fernandes et al., 2014).

Baselines. We fixed the baselines before reading any result. They are a random walk (the current log VIX), an AR(1) on log VIX estimated walk-forward, and a HAR-type regression on log VIX with 1-, 5-, and 22-day components (Corsi, 2009), also estimated walk-forward. Point accuracy uses MSE and MAE on log VIX per horizon. Distributional accuracy uses the CRPS (3) against the Gaussian predictive densities of the baselines, with residual widths estimated on TRAIN. Diebold–Mariano tests with HAC variance and the small-sample correction (Diebold and Mariano, 1995; Harvey et al., 1997; Newey and West, 1987) run per horizon and never pool horizons. The twenty-day horizon overlaps at the five-day cadence, so its loss differentials carry an overlap-corrected variance. Benjamini–Hochberg control at $q = 0.10$ applies within each table family (Benjamini and Hochberg, 1995), intervals come from the stationary bootstrap (Politis and Romano, 1994), and minimum detectable effects are printed before results. No pairwise Diebold–Mariano difference among the computed baselines is significant at any horizon or split. This design can therefore detect only large accuracy differences, and we state the limit before reporting any agentic result.

Table 3: Memory arms, TEST segment, all three horizons (runs BV1, BV2, BV3). LC slope is the learning-curve slope (4) of the rolling mean absolute standardized error per cycle, with its HAC t. MDE is the minimum detectable CRPS difference against the memory-off arm at 80% power. Corrected selection (Section 4.4); the correction changes only the five-day rows.

Arm	h	n	CRPS	cover80	PIT χ^2	LC slope (t)
Memory on	1	36	0.048	0.861	5.7	+0.001 (0.1)
	5	35	0.085	0.914	10.4	−0.010 (−4.0)
	20	32	0.129	1.000	38.6	+0.009 (1.8)
Memory off	1	37	0.048	0.838	6.5	−0.005 (−1.1)
	5	36	0.083	0.917	12.3	−0.013 (−10.6)
	20	33	0.131	1.000	20.6	+0.007 (1.8)
Shuffled	1	28	0.045	0.821	8.4	−0.001 (−0.2)
	5	28	0.074	0.929	12.7	−0.013 (−7.9)
	20	25	0.131	1.000	12.2	+0.004 (0.6)
MDE, on against off			0.0020 (h=1), 0.0041 (h=5), 0.0107 (h=20)			
Observed / MDE			0.46, 0.12, 0.15			

6.2 Results

V4 is null. Table 4 reports the evaluation for both memory-on runs. BV4 alone has rank-order informativeness at one day (Spearman 0.48, $p = 0.01$) and a positive slope at twenty days (0.73, HAC $t = 1.9$), and a single-run study would read these results as suggestive evidence. Three facts exclude that reading. First, BV1 has neither result at one day (slope −0.58, Spearman 0.12). Second, sample composition does not explain the disagreement. On the 27 dates the two runs share, BV1 gives a Spearman correlation of 0.17 ($p = 0.39$) and BV4 gives 0.51 ($p = 0.006$) against the same realizations. Third, the issued dispersions of the two runs are uncorrelated at every horizon (−0.06). Two mutually uncorrelated series cannot both measure the same property of the market. At five days the corrected series gives a slope of +0.27 ($t = 0.45$) and a rank correlation of 0.06 in BV4, and −0.05 and −0.06 in BV1. The robustness selection gives a slope of +0.31 ($t = 0.51$) in BV4. The harness selection reported −0.02 at five days in BV4, computed on the doubled values. The issued variance level never improves on the classical baselines in QLIKE at any horizon, so the loss condition of V4 fails as well.

Calibration. Figure 4 reports coverage and PIT uniformity per horizon and per split. The issued intervals are conservative everywhere, most strongly at twenty days, where TEST coverage is 1.000 in all four runs against a nominal 0.80. Together with the mean rules of Section 4.4, this pattern points to the projection layer. If the five-day width is already conservative, a width scaled by $\sqrt{20/5} = 2$ from a five-day anchor is too wide at twenty days by construction.

V5 is null. Table 5 reports the full six-class family. Five-day index moves are 31% larger after macroeconomic data days (permutation $p = 0.0075$, circular shift $p = 0.022$), and

Table 4: Dispersion informativeness (V4), TEST segment, both memory-on runs. Slope and R^2 come from regression (5) of the realized $|\Delta_h y_t|$ on the issued dispersion, Spearman is the rank correlation, and Repro. is the Pearson correlation of the issued dispersion of BV1 and BV4 on their 27 common TEST dates. The bottom rows compare the QLIKE (6) of the issued variance level with the best classical variance baseline. Corrected selection (Section 4.4), under which the three dispersions of a date are fixed multiples. The harness selection gives, for BV4 at five days, slope −0.02, Spearman 0.13, Repro. 0.72, and QLIKE 0.82 against 0.50 (DM 2.9).

Run	h	n	Slope (t_{HAC})	R^2	Spearman (p)	Repro.
BV4	1	28	+0.89 (1.8)	0.07	+0.48 (0.01)	−0.06
	5	27	+0.27 (0.5)	0.01	+0.06 (0.78)	−0.06
	20	26	+0.73 (1.9)	0.16	+0.37 (0.06)	−0.06
BV1	1	36	−0.58 (−0.5)	0.02	+0.12 (0.50)	
	5	35	−0.05 (−0.1)	0.00	−0.06 (0.74)	
	20	32	+0.41 (1.8)	0.06	+0.32 (0.07)	

QLIKE, BV4 issued against best baseline: 1.63 against 1.45 (rolling 21-day, h=1), 0.59 against 0.50 (HAR, h=5), 0.26 against 0.23 (HAR, h=20); DM 1.6, 1.5, 0.9.

Table 5: News-class response (V5), five days. Market side: mean absolute five-day index move after class days against all days, with permutation and circular-shift p-values. System side: issued five-day dispersion on class decision dates against other decision dates (run BV4, corrected selection), with the number of class decision dates and a two-sided permutation p-value. Bold entries are significant at 5%. Macro data responds on the market side only, and geopolitical on the system side only.

Class	Days	|move|, class / all	Perm. / shift p	σ, on / off (n_{on})	p
Macro data	66	0.166 / 0.127	**.008 / .022**	0.205 / 0.183 (7)	.20
Monetary policy	77	0.131 / 0.127	.38 / .37	0.196 / 0.184 (8)	.46
Geopolitical	156	0.114 / 0.127	.94 / .87	**0.201 / 0.177** (27)	**.01**
Market technical	74	0.136 / 0.127	.26 / .35	0.196 / 0.184 (9)	.44
Corporate	102	0.126 / 0.127	.53 / .51	0.190 / 0.184 (13)	.69
Other	33	0.142 / 0.127	.25 / .27	0.181 / 0.186 (7)	.79

Benjamini–Hochberg at $q = 0.10$ retains macro data within the five-day market family. Under the corrected selection the issued five-day dispersion does not widen on macroeconomic data days (0.205 against 0.183, $p = 0.20$), and the robustness selection agrees (0.217 against 0.185, $p = 0.07$). The harness selection reported 0.295 against 0.218 ($p = 0.03$). Three of the seven macro-data decision dates are twenty-day cycle dates, and on those dates the harness five-day dispersion is the unscaled twenty-day value. BV1 has the same pattern ($p = 0.05$ under the harness selection and 0.46 under the corrected selection). No class is positive on both sides at any horizon, and V5 returns null.

The system responds to a different class. Geopolitical days widen the issued dispersion at one, five, and twenty days ($p = 0.014$, 0.0135, and 0.019), and geopolitical is the one class that Benjamini–Hochberg retains within the five-day system family. The index does not move on

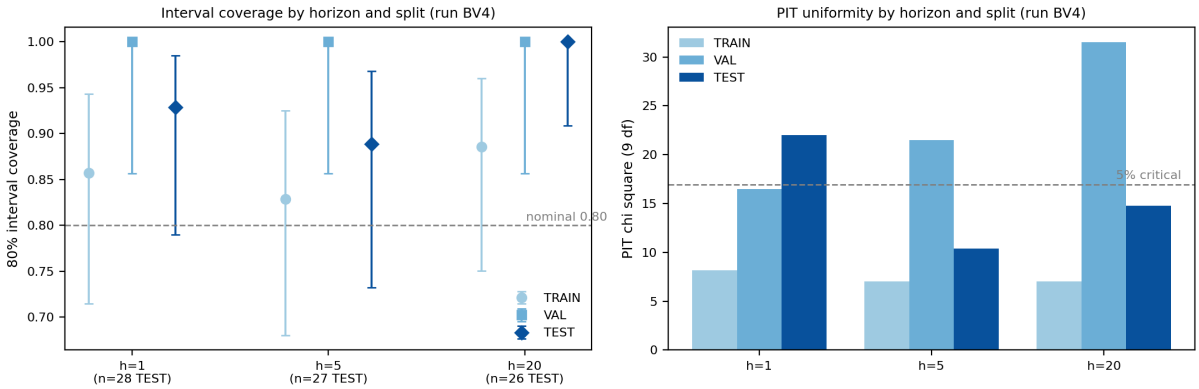

Figure 4: Calibration of the issued distributions per horizon and per split (run BV4). The left panel plots coverage (2) with Jeffreys 95% intervals. The right panel plots the PIT χ^2 statistics against the 5% critical value of the $\chi^2(9)$ reference. Coverage exceeds the nominal 0.80 at every horizon and reaches 1.000 at twenty days on TEST. Corrected selection (Section 4.4).

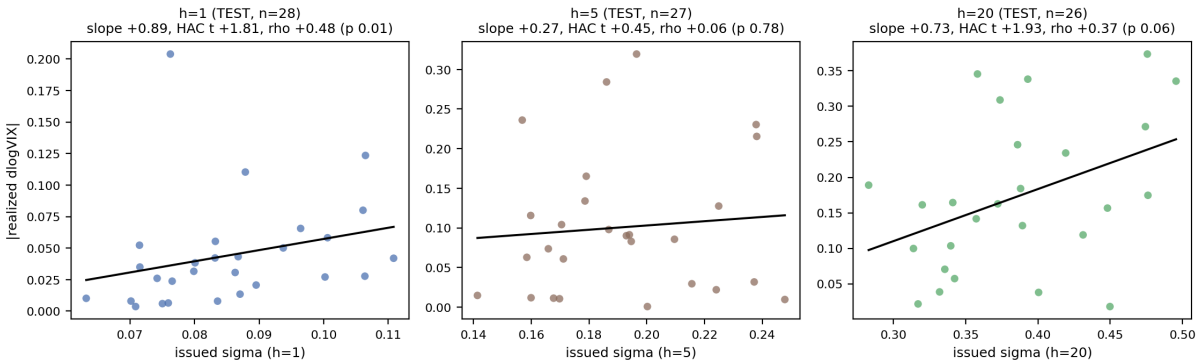

Figure 5: Issued dispersion against the realized absolute change in log VIX on the TEST segment, per horizon (run BV4), with the fitted line and the slope, HAC t, and rank correlation of each panel. Corrected selection (Section 4.4). The last column of Table 4 reports the cross-run reproducibility of each series.

those days ($p = 0.65$, 0.94, and 0.74). The system therefore widens its dispersion on headlines that carry no realized variability and does not widen it on the releases that move the index. The conjunction criterion exists to exclude a one-sided response of this kind, and for this reason we did not test the two sides separately. The family-wise conjunction test gives $p = 0.012$ under the corrected selection (0.003 under the harness selection, 0.002 under the robustness selection). Its value stays small because the market statistic of macro data is rarely attained under a common shift of the calendars. The test therefore measures the market side and cannot establish the system side alone. Five days is the only horizon whose evaluation windows do not overlap at the five-day decision cadence, which fixes the horizon of the test. One retrieval result is disclosed as well. The grounding-first verifier cites the quantitatively dense context notes more often than headline text, so the class-response measure uses class-day exposure in place of citation counts.

7 Conclusion

The protocol is the component we propose for reuse. It has five rules. Fix the criteria before the run. Give every learning claim a placebo arm and a memoryless baseline on the same calendar. Report coverage over a population whose composition is decomposed. Publish the branch-removal census with attribution. State the minimum detectable effect next to every null. Applied here, the protocol returned one clean pass (traceability), one explained null (the learning signal is the calendar), and two nulls that only the logs could establish. The issued dispersion does not reproduce across runs. The system widens its dispersion on geopolitical headlines that do not move the index, and it does not widen it on the macroeconomic releases that do.

The defect behind the last two results has the widest implication. It lies outside the language model, in projection arithmetic that leaves the five-day values of twenty-day cycles unscaled, combined with an evaluation rule that keeps the last row of a date. Under that rule the evaluation reported a reproducible searched horizon (0.72) and a system-side response to macroeconomic data ($p = 0.03$), and both results are products of the defect. A return-based evaluation would not have surfaced it. A log complete enough to recover the derivation rules and the cycle types, together with a second run for comparison, surfaced it. Two extensions follow directly. The first issues each horizon from its own search, records one forecast per date and horizon, and tests V4 and V5 again. The second replaces the scripted supervisor with practicing portfolio managers. V3 currently establishes the mechanism only, and the human factor remains untested.

8 Limitations, Ethics, and Disclosure

The study covers one index, one 25-month window, and one or two volatility regimes. The forecast comparison is a verification exercise with wide minimum detectable effects, and nothing here estimates economic value. The twenty-day horizon overlaps at the decision cadence, so its inference is weaker, and the one-day horizon is a projection in every cycle. The memory control arms run with a single seed and are labeled as such, and Table 3 states the cost in detectable effect. The system side of V5 rests on 7 macro-data and 27 geopolitical decision dates. The corrected selection applies the rules of the five-day cycles to twenty-day cycles, and the robustness selection, which drops the twenty-day cycles, gives the same verdicts. The per-class headline thresholds that define event days (2 to 5 headlines, set per class by a top-quartile rule) are a design choice, which we disclose with the thresholds in the replication package. The supervisor is programmatic throughout. TRAIN overlaps the pretraining corpus of the generator, and VAL and TEST postdate every pinned model by at least nine months. News tagged with the VIX symbol is sparse, so the news basket is built from volatility topics.

The system is a research prototype. It produces no investment advice and, by construction, issues no trades in this paper. The intended user is a professional fiduciary who retains control through a blocking interrupt for every action outside a bounded autonomy band, and the audit

trail preserves accountability for every action. Market data were obtained under the terms of service of the providers. Portions of the implementation code were developed with the assistance of an AI coding tool, and all design decisions, evaluations, and conclusions are the author's own. The replication package (notebook, configuration, run databases, figure-generation scripts, and provenance ledger) is available from the author on request.

References

Anthropic. Introducing the model context protocol. `https://www.anthropic.com/news/model-context-protocol`, 2024. Specification: `https://modelcontextprotocol.io`.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. Supervisory guidance on model risk management (SR 11-7). Technical Report SR 11-7 / OCC 2011-12, Federal Reserve Board and OCC, 2011. URL `https://www.federalreserve.gov/boarddocs/srletters/2011/sr1107.htm`.

Shuaiyu Chen, T. Clifton Green, Huseyin Gulen, and Dexin Zhou. What does ChatGPT make of historical stock returns? extrapolation and miscalibration in LLM stock return forecasts, 2024. arXiv preprint arXiv:2409.11540.

Yanxu Chen, Zijun Yao, Yantao Liu, Amy Xin, Jin Ye, Jianing Yu, Lei Hou, and Juanzi Li. StockBench: Can LLM agents trade stocks profitably in real-world markets?, 2025. arXiv preprint arXiv:2510.02209.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. arXiv preprint arXiv:2110.14168.

Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009. doi: 10.1093/jjfinec/nbp001.

Antoine Didisheim, Martina Fraschini, and Luciano Somoza. AI's predictable memory in financial analysis. *Economics Letters*, 256:112602, 2025. doi: 10.1016/j.econlet.2025.112602.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Louis H. Ederington and Jae Ha Lee. How markets process information: News releases and volatility. *The Journal of Finance*, 48(4):1161–1191, 1993. doi: 10.1111/j.1540-6261.1993.tb04750.x.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 2024.

Marcelo Fernandes, Marcelo C. Medeiros, and Marcel Scharth. Modeling and predicting the CBOE market volatility index. *Journal of Banking and Finance*, 40:1–10, 2014. doi: 10.1016/j.jbankfin.2013.11.004.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis, 2023. arXiv preprint arXiv:2309.17322.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011. doi: 10.3982/ECTA5771.

David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997. doi: 10.1016/S0169-2070(96)00719-4.

LangChain, Inc. LangGraph: Low-level orchestration framework for building stateful agents. `https://github.com/langchain-ai/langgraph`, 2024.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K. P. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, and Qianqian Xie. InvestorBench: A benchmark for financial decision-making tasks with LLM-based agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2509–2525, Vienna, Austria, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.126.

Xiangyu Li, YaWen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. HedgeAgents: A balanced-aware multi-agent financial trading system. In *Companion Proceedings of the ACM on Web Conference 2025 (WWW '25 Companion)*, pages 296–305, New York, NY, USA, 2025b. ACM. doi: 10.1145/3701716.3715232.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust LLMs' economic forecasts?, 2025. arXiv preprint arXiv:2504.14765.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 220–229. ACM, 2019.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, NIST, 2023.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.

Andrew J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, 2011. doi: 10.1016/j.jeconom.2010.03.034.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Suproteem K. Sarkar and Keyon Vafa. Lookahead bias in pretrained language models, 2024. SSRN working paper 4754678; ICML 2025 DIG-BUGS Workshop.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Robert E. Whaley. The investor fear gauge. *The Journal of Portfolio Management*, 26(3):12–17, 2000. doi: 10.3905/jpm.2000.319728.

Yihan Xia, Panpan You, Taotao Wang, Fang Liu, Han Qi, Xiaoxiao Wu, and Shengli Zhang. Agentic trading: When LLM agents meet financial markets, 2026. arXiv preprint arXiv:2605.19337.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. TradingAgents: Multi-agents LLM financial trading framework, 2024. Oral presentation, Multi-Agent AI in the Real World Workshop at AAAI 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Yuxuan Ye, Jun Han, Ao Hu, Juncheng Bu, Yiyi Chen, Liangjian Wen, Danilo Mandic, Danny Dongning Sun, Xu Yinghui, and Zenglin Xu. The alpha illusion: Reported alpha from LLM trading agents should not be treated as deployment evidence, 2026. arXiv preprint arXiv:2605.16895.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Suchow, and Khaldoun Khashanah. FinMem: A performance-enhanced LLM trading agent with layered memory and character design. *Proceedings of the AAAI Symposium Series*, 3(1):595–597, 2024a. doi: 10.1609/aaaiss.v3i1.31290.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024b.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 4314–4325, New York, NY, USA, 2024. ACM. doi: 10.1145/3637528.3671801.

Taojie Zhu, Wentao Zhao, Rui Sun, Beidi Luan, Jiacheng Lu, Sinuo Wang, Jing Li, Daxin Jiang, Yonghong He, and Zuo Bai. From knowing to doing: A memory-controlled benchmark for LLM trading agents on stock markets, 2026. arXiv preprint arXiv:2605.28359.